Towards Building the LEMI Readability Platform for Children's Literature in the Romanian Language

Madalina Chitez[1], Mihai Dascalu[2], Aura Cristina Udrea[2], Cosmin Striletchi[3], Karla Csürös[1], Roxana Rogobete[1], Alexandru Oravițan[1]

[1] West University of Timisoara, Romania, [2] National University of Science and Technology POLITEHNICA Bucharest, Romania, [3] Technical University of Cluj-Napoca, Romania

{madalina.chitez, karla.csuros, roxana.rogobete, alexandru.oravitan}@e-uvt.ro

{mihai.dascalu, aura.cojocaru}@upb.ro, cosmin.striletchi@com.utcluj.ro

Abstract

Readability is a crucial characteristic of texts, greatly influencing comprehension and reading efficacy. Unfortunately, limited research is available for less-resourced languages, especially for young populations where its impact is even higher. This paper introduces a new readability tool for children's literature in the Romanian language, explicitly targeting primary school students aged 7-11. The tool consists of a digital repository of school reading texts (self-compiled corpus) and a text analysis interface that generates automatic readability reports for uploaded short texts. The methodology involves extracting, testing, and calibrating a readability formula for Romanian using the children's literature corpus. Related work on readability and readability tools is discussed, followed by a description of the children's literature corpus and the platform functionalities. The first steps are presented towards validating the readability formula for children's literature in Romanian using the ReaderBench framework, while calibration variables relevant to the Romanian language and children's literature are examined. Currently, no existing platform integrates a research-based readability formula for the Romanian language, making this tool unique. Overall, this research contributes to applied corpus linguistics and Digital Humanities studies and offers a valuable resource for educators, parents, and children in accessing age-appropriate and readable texts.

Keywords: children's literature readability platform, Romanian children's literature corpus, readability for the Romanian language

1. Introduction

Readability is a key characteristic of texts as linguistic constructs to be processed by the human mind. It refers to the set of features that influence the reader's understanding and reading efficacy (Collins-Thompson, 2014). These features include linguistic factors that have been demonstrated to have an impact on how difficult the text is perceived: lexical variation and sophistication (e.g., simple versus complex words), syntactic complexity (e.g., shorter versus longer sentences), coherence and cohesion (i.e., connection between sentences and arguments in texts) or text structure. From an educational point of view, the readability of the texts used in schools has been a major concern for numerous pedagogy studies which have looked into topics such as reading and learning motivation (Moley et al., 2011) or finding the zone of proximal development (ZPD, Vygotsky, 1978), which means using the right text level for optimal learning growth (Shanahan et al., 2016). These studies rely on rigorous linguistic research where readability is measured using readability formulas (Spache, 1953; Begeny & Greene, 2014; Lee & Lee, 2023), which have been more or less successful in assessing the level of linguistic complexity of texts. The availability of such formulas (see section 4) varies according to language and the amount of linguistic data that has been collected for that particular language. In recent years, readability formulas have been integrated into digital tools that can automatically assess text readability, among other parameters.

As part of an applied digital humanities research project, we have developed the LEMI readability platform[1] that uses our readability formula for Romanian language texts to evaluate children's literature texts for primary school students (aged 7-11). The two main functionalities of our tool are: (a) a digital repository of school reading texts based on a self-compiled corpus, where texts are distributed into readability level, and (b) a text analysis interface, which issues an automatic readability report for any uploaded short text.

This paper presents the new tool (i.e., platform) and introduces the methodology of using the school text corpus to extract, test, and calibrate the readability formula for Romanian to be integrated into the platform. The LEMI Romanian children's literature corpus[2] is made publicly available and can be used for further studies. We begin by addressing the related work concerning readability and readability tools. This is followed by the presentation of the children's literature corpus and of the platform in which main corpus-based functionalities were created. We then describe the analysis and the main results of the readability formula validation process for our platform by running the corpus through the ReaderBench framework (Dascalu et al., 2017). ReaderBench is the only existing text complexity assessment platform that includes Romanian. We end with a discussion on the calibration variables relevant to a readability formula for the Romanian language and for the analyzed type of texts (i.e., children's literature) and draw several applied linguistics conclusions.

[1] The LEMI platform is accessible at: https://lemi.ro.

[2] Github repository available at: https://github.com/chia-NR/LEMI-Romanian-children-literature-corpus

LREC-COLING 2024, pages 16450–16456

2. Related Work

2.1 Readability: Beyond Education

Researchers from various disciplines have shown a keen interest in evaluating the readability of texts. This interest emerged more than a century ago when educators in the U.S. realized that the increasing educational demand after the Great Depression was best supported by appropriate reading texts. A study by Leavy & Grey (1935) was among the first to propose a set of features that influence readability (i.e., content, style, format, structure), thus finding out that style, through sentence length and word complexity, is the most influential. This was followed by a series of research initiatives designed to capture the readability level of a text through formulas (see below). Since their emergence in the 1950s, the core concepts behind readability formulas have remained stable. As DuBay (2004) pointed out, "by the 1980s, there were 200 formulas and over a thousand studies published on the readability formulas attesting to their strong theoretical and statistical validity" (p. 2). Metrics such as word length, sentence length, or lexical choices are primary indicators. It is generally understood that texts with shorter sentences and more common words are easier to understand than their longer and more complex counterparts.

The number of methods to assess readability has grown considerably, and new models tailored for various disciplines are regularly introduced. These models have proved their effectiveness for a multitude of sectors: for education (e.g., textbook content) as well as for all healthcare, law, business, public administration, or research areas investigating accessibility – i.e., readability of written documents for specific audiences or the wider public (DuBay, 2004).

2.2 Readability Formulas

The most commonly used readability formulas have been first developed by U.S. researchers, so they have been created based on the analysis of the English language and texts:

- In the 1940s: the Dale-Chall formula (Dale & Chall, 1948), including sentence length and list of familiar words; the Flesch Reading Ease (FRE; Flesch, 1948) used sentence length and syllable count.
- In the 1950s: the Gunning Fog Index (GFI; Gunning, 1952) used sentence length and the percentage of complex words;
- In the 1970s: the Fry Readability Graph (FRG; Fry, 1968) focused on sentence length and syllables per 100 words; the Simple Measure of Gobbledygook (SMOG; McLaughlin, 1969); the Flesch-Kincaid Grade Level Readability Test (F-K; Kincaid et al., 1975);
- Other relevant indices: Coleman-Liau index (CLI) (Coleman & Liau, 1975) and Automated Readability Index (ARI) (Smith & Senter, 1967) take into account characters instead of syllables per word.

In the 1960s, formulas for languages other than English started to be created. For example, *Läsbarhetsindex* (LIX) was developed in Sweden by Carl-Hugo Björnsson (1968) in the late 1960s, and it has since been used for eleven languages: Swedish, Norwegian, Danish, English, French, German, Italian, Spanish, Portuguese, Finnish, Russian (Björnsson, 1983). While LIX considered word and sentence length, much like the Flesch test, it uniquely considered the number of long words in a text instead of syllable count. Other formulas have been tested for each language:

French	Kandel and Moles Index	Kandel & Moles, 1958
Italian	Gulpease Index	Lucisano & Piemontese, 1988
Spanish	INFLESZ scale	Barrio-Cantalejo, 2008

Table 1: Readability formulas for other languages

In the past twenty years, research has expanded considerably, the list of readability formulas being nearly endless nowadays (Stellner, 2013, p. 24).
As for readability formulas for Romanian, studies have been relatively scarce. Two studies by Garais (2011) and Garais & Enaceanu (2011) proposed a readability formula based on standard L1 and L2 formulas, where an L1 formula (e.g., FRE) is a formula resulting in a 0 to 100 scale and an L2 formula (e.g., SMOG) indicates the level of necessary education to understand the text. Their formula included text length measured in the number of characters. Several other papers evaluate complexity features for texts written in Romanian with the help of the ReaderBench framework, which was used for automated writing evaluation (Sirbu et al., 2018) or automated essay scoring (Toma et al., 2021) studies.

2.3 Tool Integration

As of today, there is no platform, tool, or app that integrates a research-based readability formula for the Romanian language. The same is valid for digital instruments that offer access to children's literature texts based on readability levels or that can assess the readability of given texts automatically. Similar tools to what we have developed exist, however, for English: Text Inspector, developed in the U.K. (Bax, 2012), and ARTE, developed in the U.S. (Choi & Crossley, 2021). Our tool also offers access to a self-compiled digital repository of children's texts, which can be filtered by different criteria (e.g., age, grade, readability level), when compared to the previous two systems.

3. Method

3.1 Corpus

We began working on two key undertakings to create a representative corpus for our digital repository. First, we analyzed the electronic versions of the Romanian language and literature textbooks approved by the

Ministry of Education for primary school (grades I-IV), which are publicly available at manuale.edu.ro. This was done in order to create a database of all reading texts that appear in primary textbooks, as they are the texts that children interact with most frequently. Then, we sent a series of print and online surveys to three target groups (grades I-IV): teachers, parents, and children. All target groups were asked to (1) evaluate their satisfaction with their Romanian textbooks; (2) give examples of texts that children at their level enjoy; (3) give level-appropriate recommendations of authors and/or texts.

We compiled a list of reading texts readily accessible through the digital repository for teachers, parents, and children, by combining the results from the textbook analysis and the surveys. The texts are either original or adapted to suit a particular level. The featured texts are primarily fiction or poetry, written by both Romanian and foreign writers. We also took into consideration the publication date of the texts, marking them as either classic or modern. For this study, we have selected a subset of 80 texts, 20 for each grade level; 56.25% (45) were written by classic Romanian writers, 11.25% (9) by modern Romanian writers, 11.25% (9) by classic foreign writers, and 21.25% (17) by modern foreign writers. Other relevant aspects considered in our metadata include domains (e.g., arts, geography, linguistics, advice) and themes (e.g., animals, adventure, childhood, drama, family, fantasy, history, nature, science fiction, humor) – see Figure 1 for a sample text.

<RD1_RO_C_033>

Odată, vara, ies din casă. Merg la unchiul Vasile să fur niște cireșe.
Mă prefac că îl caut pe vărul Ioan.
Mă ascund în cireș și mănânc. Mătușa vine și mă vede. Eu fug iute prin grădină. Ea mă urmărește, dar nu mă prinde. Ajung acasă repede și stau cuminte.
Spre seară, unchiul Vasile, paznicul și un alt om vin la tata acasă. Sunt supărați din cauza cireșelor pe care le-am furat. Tata plătește pentru pagubă. Apoi, tata mă ceartă pentru ce am făcut. Eu gândesc: „Deși mă feresc să fac prostii, parcă mereu se întâmplă ceva".

Figure 1: Sample text in children's literature corpus

We have also assigned appropriate age groups to each level (e.g., 7-8 years old for grade I, 8-9 years old for grade II) and a three-level reading complexity mark for each grade to distinguish between easier and harder to read texts meant for the same age group. For this study, we used a corpus (i.e. LEMI Romanian children's literature corpus) consisting of 33,154 words. The word frequency indicates a notable prevalence of pronouns, alongside recurrent instances of familial terms such as 'mother' and 'father,' as well as the concept of 'home.' Moreover, there is a significant occurrence of verbs representing elementary actions, such as 'to see,' 'to say,' 'to do,' 'to come,' and 'to hear'. The median syllable count in the corpus is 2.

3.2 Readability Platform for Romanian

The LEMI platform is a cross-platform web-based application that categorizes its users into three classes. The software areas accessible by each user category offer specific functionalities.

Visitors can view a series of generic information (project description and credentials, tutorials, contact information) and have the account creation option. After filling in the required information, the visitors become registered users who can authenticate themselves using their credentials, and the main functionalities become accessible.

Authenticated users can benefit from browsing the entire text collection. The texts can be filtered by (a) domain and theme: by selecting the desired items from a list of available options, the user can narrow the category of the displayed information; (b) keywords: specifying searched words and even sentence parts allows the user to isolate the texts that contain the searched information. The * wildcards are supported and can replace word parts and/or entire words, making the searching process very effective. After filtering, the resulting text list is displayed with the associated metadata (title, author, domain, theme, age groups to whom the texts are addressed, reader class, and complexity). The textual information is reduced to an excerpt with a predefined character length to make the text list more compact. If the search was performed using keywords, the searched information is highlighted in the displayed text portion. The user has the option of reading the entire text online. The texts are also downloadable in PDF format (see Appendices A-C). Another option accessible to authenticated users involves analyzing the complexity of their own uploaded texts. By selecting the files to be analyzed, the user can have the readability formula applied to their material, the result being displayed in recommended age and / or standard school grade, linguistic complexity levels, and text overall intricacy. This valuable functionality allows educators to ensure that their materials are suitable for the study formations.

Administrators have access to a secured Control Management System (CMS) to manage the entire database. Administrators control the texts composing our corpus and the associated metadata. The platform offers CRUD facilities (Create, Read, Update, Delete). Multicriterial searching and filtering are implemented. Moreover, administrators can view the list of registered users and the texts uploaded by each user can be also inspected from the administration interface, if consent is provided.

The connection between the main functionalities of the LEMI platform and the field of Digital Humanities is reinforced by three main characteristics: (a) the access to the first digital repository of children's literature texts in Romanian; (b) the literary heritage aspect of LEMI is enhanced by the inclusion in the digital repository of not only the most popular pieces of children's literature (national and international) but also of literary samples of texts that have not been available to the general public in user-friendly format (texts from old Romanian textbooks, pieces of

literature by authors representing the language minorities of Romania, rarely used pieces of literature by classic Romanian authors); (c) the computational interface in LEMI reflects latest developments in Digital Humanities research, where linguistics and computational linguistics intersect (Luhmann & Burghard, 2022).

4. Results and Discussion

4.1 Corpus Analysis with ReaderBench

ReaderBench (Dascalu et al., 2013) is a useful tool for conducting a textual analysis, as it provides a comprehensive list of textual complexity indices, such as part-of-speech extraction, syntactic dependencies, coherence evaluation between sentences, paragraphs, or within a paragraph, word statistics, and exploration of polysemous words.

We performed an evaluation of textual intricacy by employing ReaderBench's complexity indices on the corpus composed of Romanian texts. We identified key discriminating indices displayed in Table 2 through the usage of the strategy involving the selection of *k* best features and a correlation matrix to reduce the number of highly correlated features.

Feature	Description	Importance
Max(Dep_xcomp / Par)	Maximum number of 'xcomp' (open clausal complements) dependencies per paragraph	0.266
SD(ParseDepth / Sent)	The standard deviation of the depth of the parsing tree per sentence	0.233
Max(NgramEntr_2 / Word)	Maximum entropy of bigrams per word	0.122
M(WdEntr / Par)	Mean of word entropy per paragraph	0.067
Max(UnqPOS_noun / Sent)	Maximum unique nouns per sentence	0.019

Table 2: ReaderBench complexity indices representative for the corpus

R^2 is a coefficient that displays the mutual relation between the ground truth and the prediction model (Chicco et al., 2021). A series of R^2 values were generated through Linear Regression to measure to which extent the key textual indices selected previously explain variance in the dependent variable. A value closer to 1 means the model captures the variance well.

We also used the aforementioned features in training a Random forest regressor (see Figure 2), which resulted in the following performance metrics: The Final Model – Mean Squared Error (MSE) reached 0.6163 with an R^2 of 0.4652. The resulting R^2 value suggests that a notable proportion of the variance remains unaccounted for by the model, highlighting potential areas for further refinement.

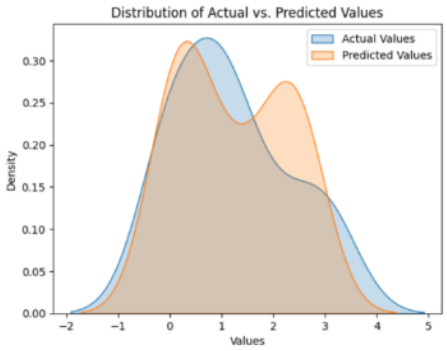


Figure 2: Random forest regression model

At this research phase, the decision to employ a large language model was not deemed practical due to the corpus's limited size. Even though language models exist for Romanian and Transformer-based models could have been trained, we opted to create an easily reproducible formula based on linguistic features. Moreover, we wanted to create a model whose predictions could be argued by teachers. However, as we anticipate the corpus to evolve and expand, the application of a larger models will become a more viable option for subsequent analyses.

4.2 Calibration of a readability formula for children's literature texts in Romanian

It should be noted that the preliminary classification of the texts (the design of the three-level reading complexity system for each grade) was based on the pedagogical perspective and experience of the research team and school partners in the project. Our analysis indicates that initial assessment and text classification per grade correlate with the results of the ReaderBench textual indices. The pedagogical evaluation took into account criteria such as content (topic), sentence length, coherence, syntactic complexity, and lexical features (e.g., frequency of long versus short words).

One of the textual complexity indices in Reader Bench that performed well in discriminating between the grades was the maximum number of bigram entropy per word. The values of this index are shown in the boxplot from Figure 3, with the highest values being for 3rd and 4th grade, the smallest one for 1st grade, and moderate values were observed for the 2nd grade. A high value indicates that the text has a wide variety of word combinations, making it more complex, while a low value means more repetitive or predictable word combinations.

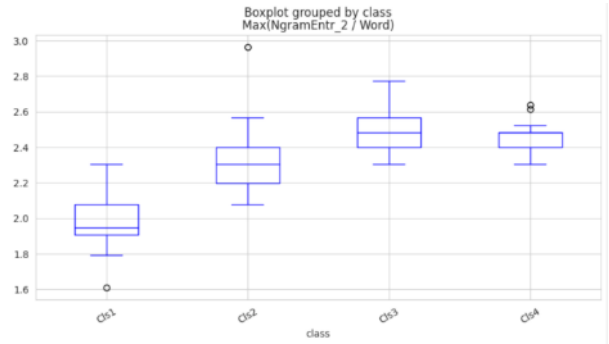


Figure 3: Bigram entropy per word

In order to further validate the results and to calibrate the variables relevant to a readability formula for the Romanian language, the next step in the research is to evaluate the ReaderBench text classification in class. A tentative 'user calibration' carried out in October 2023 in schools for grades 3 and 4 (256 children participated in the study) resulted in an acceptance rate of text-per-level of 96.48%. This process will be replicated for a larger set of texts and delivered to all school grades in the project (0-4). This multi-step validation process will enable a selection of texts that align with the cognitive and linguistic abilities of the targeted age group and design the tool that will automatically assess user-uploaded texts.

5. Conclusions

The importance of creating readability-related language applications for educational purposes is incontestable. Doing that for less-resourced languages, such as Romanian, is a complex process that requires linguistic data collection, appropriate analysis with available resources, and didactic validation. In this paper, we have presented several major steps that precede launching the first version of the readability platform for children's literature in Romanian, planned for November 2023. We have created a repository of children's texts distributed on readability levels (from grade 1 to 4) and the readability platform. The corpus analysis in this paper with the ReadearBench text complexity framework indicates a satisfactory match between texts in the corpus and relevant metrics in the framework (e.g., bigrams). This, together with further classroom validation stages, will help us calibrate the final version of the readability formula to be integrated into the platform for automatic text evaluation of uploaded short texts. Our research is expected to impact both the corpus linguistics and Digital Humanities areas in Romania, as the platform will have a significant educational impact.

6. Ethical issues

The research conducted within the project involved the processing of a large amount of data collected from schoolchildren (grades 0-4) and was based on the voluntary participation of research subjects – schoolchildren, their parents (or legal tutors) and their instructors. The questionnaires collected information regarding the pupils' class level, location of the school (urban vs rural), textbook used for Romanian classes, opinion related to the quality of the texts included in textbooks, types of texts included in the coursebooks or desired in such textbooks, favorite authors and texts, other useful reading materials, opinion regarding the utilization of digital platforms in selecting reading texts.

All the collected metadata were available exclusively to research project members. Data were processed for statistical and scientific research purposes only, without any alteration from the research team. The analysis was based on the informants' fill-in-the-blank input and was uniformized by the research team. All participants (schoolchildren, parents, and instructors) were previously informed about the key elements of the research study and what their participation will entail. At the beginning of the project, a partnership agreement was signed with the educational institutions (secondary education institutions) where the study was carried out. The informed-consent papers consisted of a written consent document containing: (a) a summary of the project – purpose and objectives, duration, host institution, contact person; (b) the data collection procedure (data collection, data anonymization, data storage on web application); (c) details on the possibility of withdrawal; (d) the expected benefits and results: access to pedagogical recommendations and data statistics; (e) a GDPR section and a declaration of consent and personal signature for the use of the delivered text or survey within the project. The documents were adapted to the level/category and readability level of prospective participants, in order to enhance participants' understanding. Protection of all personal data was assured following GDPR regulations.

All children's literature texts which have been included in the corpus and which are going to be made freely available via the readability platform comply with the Romanian and international legislation in point of copyright. We have made sure that all categories of texts (Romanian classical and modern literature as well as Romanian translations of classical and modern international literature) do not infringe any copyright. Most texts are short fragments (2-3 pages) of larger literary work, which complies with Article 35, paragraph (d) from the Copyright Law No.8 (14 March 1996). Numerous texts have been adapted (simplified versions of original text, lexical items replaced), and some have been translated into Romanian by the research team.

7. References

7.1 Bibliographical References

Barrio-Cantalejo IM, Simón-Lorda P, Melguizo M, et al. [Validation of the INFLESZ scale to evaluate readability of texts aimed at the patient]. Anales del Sistema Sanitario de Navarra. 2008 May-Aug; 31(2): 135-152. DOI: 10.4321/s1137-66272008000300004. PMID: 18953362.

Bax, S. (2012). *Text inspector. Online text analysis tool*. Available at: https://textinspector.com/.

Begeny, J. C., & Greene, D. J. (2014). Can readability formulas be used to successfully gauge difficulty of reading materials?. *Psychology in the Schools*, *51*(2): 198-215.

Björnsson, C. H. (1968). *Läsbarhet*. Liber, Stockholm.

Björnsson, C. H. (1983). Readability of newspapers in 11 languages. *Reading Research Quarterly*: 480-497.

Chicco, D., Warrens, M.J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7: e623. doi: 10.7717/peerj-cs.623. PMID: 34307865; PMCID: PMC8279135.

Choi, J.S. and Crossley, S. (2021). *Readability Assessment Tool for English Texts* (No. 6190). EasyChair.

Coleman, M. and Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2): 283–284. https://doi.org/10.1037/h0076540.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, *165*(2): 97-135.

Dale, E. and Chall, J.S. (1948) A Formula for Predicting Readability. *Educational Research Bulletin 27*(1): 11-28.

Dascalu, M., Dessus, P., Trausan-Matu, S ̧., Bianco, M., and Nardy, A. (2013). Readerbench, an environment for analyzing text complexity and reading strategies. In *Artificial Intelligence in Education: 16th International Conference, AIED* 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16, Springer, pp. 379-388.

Dascalu, M., Gutu, G., Ruseti, S., Paraschiv, I. C., Dessus, P., McNamara, D. S., Crossley, S.A., & Trausan-Matu, S (2017). *ReaderBench*: A Multi-lingual Framework for Analyzing Text Complexity. In E. Lavoué., H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds), *Data Driven Approaches in Digital Education.* EC-TEL 2017. Lecture Notes in Computer Science, vol 10474. Springer, Cham. https://doi.org/10.1007/978-3-319-66610-5_48

DuBay, W. H. (2004). *The principles of readability*. *Online Submission*. Available at: https://files.eric.ed.gov/fulltext/ED490073.pdf.

Flesch, R. (1948). A new readability yardstick*, Journal of Applied Psychology*, 32(3): 221-233. doi: 10.1037/h0057532

Fry, E. (1968). A readability formula that saves time. *Journal of reading,* 11(7): 513-578.

Garais, E. G. (2011). Web applications readability. *Romanian Economic Business Review* 5: 117-121.

Garais, G. E., & Enaceanu, A. S. (2011). Determining quality levels for improving maintenance processes. *Annals of DAAAM & Proceedings*.

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill.

Kandel, L. and Moles, A. (1958). Application de l'Indice de Flesch à la langue française*. Cahiers d'Etudes de Radio-Television 19*: 253-274.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Research Branch Report*.

Leary, B. E., & Gray, W. S. (1935). *What Makes a Book Readable: With Special Reference to Adults of Limited Reading Ability...* University of Chicago Press.

Lee, B. W., & Lee, J. H. J. (2023). Traditional readability formulas compared for English. *arXiv preprint arXiv:2301.02975*.

Lucisano, P., & Piemontese, M. E. (1988). Gulpease: una formula per la predizione della leggibilita di testi in lingua italiana. *Scuola e città 3*: 110-124.

Luhmann, J., & Burghardt, M. (2022). Digital humanities - A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape. *Journal of the Association for Information Science and Technology*, *73*(2), 148-171.

McLaughlin, G. H. (1969). SMOG grading: A new read-ability formula. *Journal of Reading*, 12(8): 639-646.

Moley, P. F., Bandré, P. E., & George, J. E. (2011). Moving beyond readability: Considering choice, motivation and learner engagement. *Theory into Practice*, *50*(3): 247-253.

Shanahan, T., Fisher, D., & Frey, N. (2016). The challenge of challenging text. In: Scherer, M. (Ed.). (2016). *On developing readers: readings from educational leadership (EL Essentials)* (pp. 100-109). Alexandria, USA: ASCD.

Sirbu, M. D., Botarleanu, R. M., Dascalu, M., Crossley, S. A., & Trausan-Matu, S. (2018). ReadME–Enhancing Automated Writing Evaluation. In *Artificial Intelligence: Methodology, Systems, and Applications: 18th International Conference*, AIMSA 2018, Varna, Bulgaria, September 12–14, 2018, Proceedings 18 (pp. 281-285). Springer International Publishing.

Smith, E. A., & Senter, R. J. (1967). *Automated readability index* (Vol. 66, No. 220). Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command. Wright-Patterson Air Force Base, Ohio.

Spache, G. (1953). A new readability formula for primary-grade reading materials. *The Elementary School Journal*, *53*(7): 410-413.

Stellner, B. (2013). *Readability of Quarterly Reports: Do Companies Mislead Investors*?. Anchor Academic Publishing (aap_verlag).

Toma, I., Marica, A.-M., Dascalu, M., & Trausan-Matu, S. (2021). ReaderBench – Automated

Feedback Generation for Essays in Romanian. *Scientific Bulletin*, University Politehnica of Bucharest, Series C, 83: 21-34.

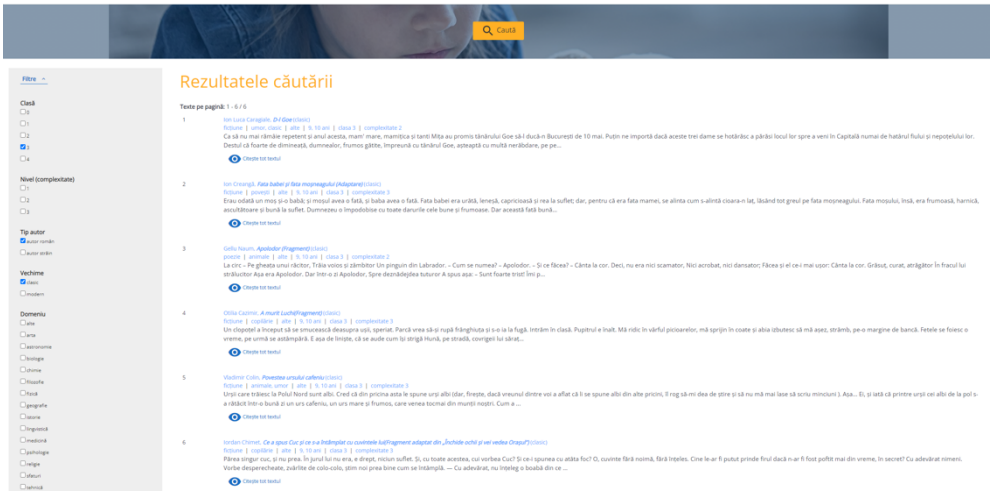
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

7.2 Language Resource References

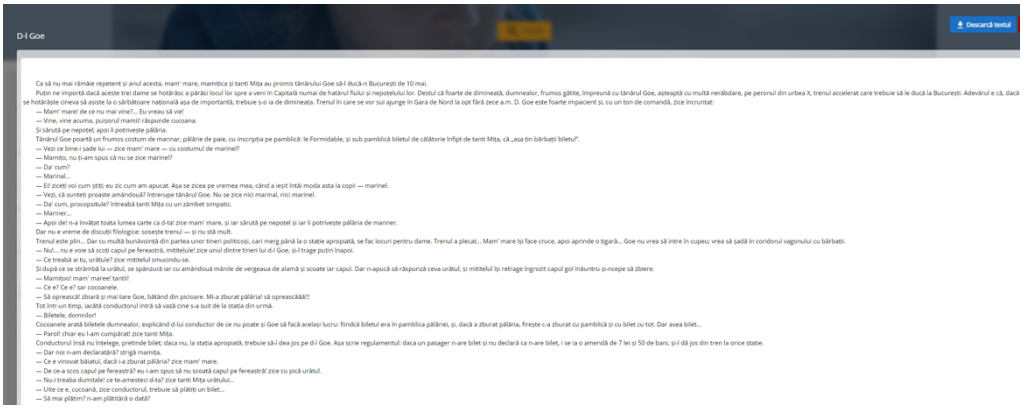
LEMI Romanian children's literature corpus is publicly available on GitHUb (https://github.com/chia-AR/LEMI-Romanian-children-literature-corpus).

7.3 Appendices

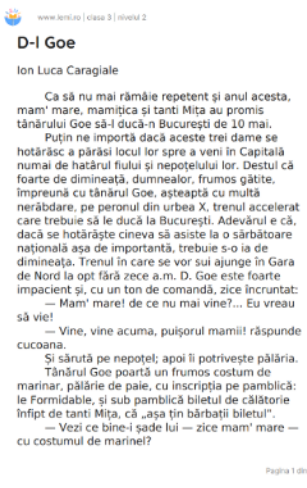
Appendix A. Screenshot with the search functionality within the LEMI repository



Appendix B. Screenshot of selected text (from search list) to be read online



Appendix C. Screenshot of downloadable format (customized PDF) for selected text (from search list)



D-l Goe

Ion Luca Caragiale

Ca să nu mai rămâie repetent și anul acesta, mam' mare, mamitica și tanti Mița au promis tânărului Goe să-l ducă-n București de 10 mai.

Puțin ne importă dacă aceste trei dame se hotărăsc a părăsi locul lor spre a veni în Capitală numai de hatârul fiului și nepoțelului lor. Destul că foarte de dimineață, dumnealor, frumos gătite, împreună cu tânărul Goe, așteaptă cu multă nerăbdare, pe peronul din urbea X, trenul accelerat care trebuie să le ducă la București. Adevărul e că, dacă se hotărăște cineva să asiste la o sărbătoare națională așa de importantă, trebuie s-o ia de dimineață. Trenul în care se vor sui ajunge în Gara de Nord la opt fără zece a.m. D. Goe este foarte impacient și, cu un ton de comandă, zice încruntat:

— Mam' mare! de ce nu mai vine?... Eu vreau să vie!

— Vine, vine acuma, puișorul mamii! răspunde cucoana.

Și sărută pe nepoțel; apoi îi potrivește pălăria.

Tânărul Goe poartă un frumos costum de marinar, pălărie de paie, cu inscripția pe pamblică: le Formidable, și sub pamblică biletul de călătorie înfipt de tanti Mița, că „așa țin bărbații biletul".

— Vezi ce bine-i șade lui — zice mam' mare — cu costumul de marinel?